AD-A032 091   RAND CORP   SANTA MONICA CALIF   F/G 5/9
DISCRIMINANT ANALYSIS AND CLASSIFICATION OF TEACHING EFFECTIVEN--ETC(U)
MAR 76   J E WARE, R G WILLIAMS

UNCLASSIFIED   P-5600   NL

1 OF 1
ADA032091

END
DATE
FILMED
1 - 77

(2)

B S.

(6)

DISCRIMINANT ANALYSIS AND CLASSIFICATION OF TEACHING EFFECTIVENESS USING STUDENT RATINGS: THE SEARCH FOR DOCTOR FOX

(10)

John E. Ware, Jr. and Reed G. Williams

(11) March 1976

(12) 21p.

(14) P-5600

296 600

VB

ACKNOWLEDGMENTS

Data gathering and analysis was supported by the Office of Research and Projects, Southern Illinois University at Carbondale. Students were scheduled with the cooperation of Larry Bush and John Mouw, Southern Illinois University at Carbondale. The preparation of this paper was supported by The Rand Corporation, Santa Monica, California, 90406.

A summary of this paper was presented at the Fourteenth Annual Conference on Research in Medical Education, November 1975, during the 86th Annual Meeting of the Association of American Medical Colleges, Washington, D.C.

ABSTRACT

This investigation employed discriminant analysis to improve the usefulness of student-faculty ratings in detecting differences in lecturer types. Equivalent groups of college students viewed lectures delivered by a Hollywood actor so as to vary in number of substantive teaching points covered (high, low) and presentation manner (enthusiastic, unenthusiastic). Students rated lecturer effectiveness using an 18-item questionnaire like those commonly used. Optimal scoring methods were derived in the first study for the purpose of differentiating among lecturer types and were cross-validated in a second study of groups of students who saw and rated the same lectures. Scoring methods derived in the first study were valid in relation to differences in faculty enthusiasm in both studies but were not valid in the second study for differences in information-giving. Results were explained in terms of the "Doctor Fox Effect" and suggestions were offered for future research.

DISCRIMINANT ANALYSIS AND CLASSIFICATION OF TEACHING EFFECTIVENESS USING STUDENT RATINGS: THE SEARCH FOR DOCTOR FOX

Inquiries into the validity of student-faculty ratings and the effects of differences in instructors on students are rarely conducted using experimental methods. Only six reports of such studies could be found. Mastin (1963) reported that high school students who heard lectures by "enthusiastic" teachers learned more and had more favorable course attitudes than students of less enthusiastic teachers. Coats and Smidchens (1966) showed that "dynamic" lectures result in higher student achievement than less dynamic lectures. Zelby (1974) demonstrated that faculty can teach so as to obtain more favorable student ratings for particular instructor and course characteristics. Powell (1975) showed that students who were required to do less work and who learned less rated their faculty more favorably than students who were required to do more work and who learned more.

The two "Doctor Fox" studies of lecture presentations (Ware and Williams, 1975; Williams and Ware, 1976) have shown that: a) an enthusiastic presentation manner results in greater student learning when initial motivation to learn is low, b) differences in information-giving produce corresponding differences in student learning levels, c) student-faculty ratings *are* valid in relation to information-giving and group learning when presentations are *not* given in an enthusiastic manner, d) the latter is *not* true when faculty presentations *are* given enthusiastically. In other words, ratings of enthusiastic presentations containing a lot of information do not differ from ratings of enthusiastic presentations containing little or no information even though students who viewed high information presentations learned more. This phenomenon, which has been termed "The Doctor Fox Effect," suggests that student ratings as commonly scored primarily reflect faculty enthusiasm (Ware and Williams, 1975).

Throughout the Doctor Fox experiments and most correlational studies of the validity of student ratings, simple methods of computing student-faculty rating scores have been employed, namely, analysis of

single-item scores or the simple algebraic sum of item scores. Such practices lag behind the scoring systems that could be developed using multivariate statistical techniques and experimental data. Easy to use computer programs required to perform multivariate analyses of student ratings (factor analysis, regression analysis, discriminant analysis, etc.) are readily available, however, the required experimental data has only recently become available and only to a limited degree.

Experiments are needed in order to precisely control faculty differences and to control for differences in student groups due to self-selection of faculty types. Precise control of differences in faculty characteristics is one way of establishing *criteria* of instructional effectiveness against which to develop and test more valid student-faculty rating scoring methods. Given such data, it is possible to develop and validate a student rating scoring algorithm which will maximally detect *known* differences in faculty characteristics under controlled conditions. These controls were achieved during the "Doctor Fox" studies by using a carefully programmed Hollywood actor who portrayed a variety of faculty types with equivalent groups of students (Ware and Williams, 1975; Williams and Ware, 1976).

The analyses presented in this paper were designed to take advantage of the data gathered during the "Doctor Fox" studies in order to determine the extent to which more sophisticated student-faculty rating scoring methods would improve the validity of rating scores in relation to differences in instructional effectiveness. Specifically, the current studies were designed to: a) derive a set of weights for student-faculty rating item scores that maximize differences in mean rating scores for groups of students known to have experienced real differences in instructional effectiveness, and b) attempt to cross-validate the new scoring methods in a second study of students who rated lectures under the same conditions.

Method

Participants

Participants in the first study were 207 undergraduate and graduate students who were enrolled in general studies sections. Thirty-three percent were males. They ranged in age from 17-42 years with a median age of approximately 20 years. Twenty-one percent were freshment, 30 percent were sophomores, 28 percent were juniors, 18 percent were seniors and 3 percent were graduate students. Forty-two percent reported they were in liberal arts, biological and physical sciences. Other academic majors included education, engineering and technology, business, and home economics. The analysis sample selected from the first study included the 115 students who experienced lectures high or low in information giving (Ware and Williams, 1975).

Participants in the second study were 213 students enrolled in 12 sections of an undergraduate psychology course. Fifty-eight percent were males. They ranged from 18 to 38 years of age with a median of approximately 20 years. Seven percent were freshmen, 47 percent sophomores, 37 percent juniors, 7 percent seniors and 2 percent were graduate students. Forty-seven percent reported they were education students, 25 percent liberal arts and sciences, 18 percent home economics, agriculture, and business. The second study analysis sample consisted of the 70 students who experienced lectures high or low in information-giving and who were not given an incentive to learn (Williams and Ware, 1976).

Faculty Characteristics. Two faculty characteristics (information-giving and presentation manner) that are frequently cited as operational definitions of teaching effectiveness were manipulated. The specific definitions and controls used are documented elsewhere and only a brief description will be repeated (Ware and Williams, 1975; Williams and Ware, 1976). High and low amounts of information-giving were achieved *vis-a-vis* strict adherence to verbatim lecture scripts during videotaped lecture presentations over the same topic (the biochemistry of learning). Teaching points were eliminated through a modified random procedure so

that the lecture high in information covered 26 teaching points and the lecture low in information covered only four teaching points.

Presentation manner was manipulated by programming the actor to give each of the two lectures (high and low information-giving) in either an enthusiastic or unenthusiastic manner. Levels of enthusiasm were associated with differences in expressiveness, vocal inflection, friendliness, charisma, humor, and "personality." The two faculty characteristics chosen for study (enthusiasm and information-giving) have been identified in previous correlational and experimental studies of faculty characteristics as associated with effective teaching (Coats and Smidchens, 1966; Rosenshine and Furst, 1971; Coffman, 1954; Isaacson, et al., 1964; Solomon, 1966).

The result was four faculty lecturer types, for purposes of the current study, as follows: (1) high information-high enthusiasm, (2) high information-low enthusiasm, (3) low information-high enthusiasm, and (4) low information-low enthusiasm.

Study Design. In the first study, sections of the same class were divided randomly. In the second study, intact sections of the same class constituted the study groups. In both studies, lecturer types were randomly assigned to student groups and the groups were shown to be equivalent in terms of age, sex, GPA, and a priori interest in the lecture topic. Students in all groups: a) viewed one lecture presentation, b) rated the presentation using an 18-item questionnaire like those in general use (Pohlmann, 1975), and c) were tested over the material (test score based on 26 multiple-choice questions). The 18 rating items, which were scored using a five-choice response continuum, are listed in Figure 1.

Analysis Plan. Discriminant Analysis (Tatsuoka, 1971, Huberty, 1975) and student ratings were used to solve for the linear discriminant functions (LDF's) that maximize differences among means for groups of students in the first study. In other words, the following question was asked: How should student-faculty ratings be scored in order to reduce the amount of overlap in mean rating scores of students experiencing known differences in teaching effectiveness (information-giving and presentation manner)? Eighteen rating variables were used in a

stepwise manner in solving for functions discriminating among the four groups in the first study. LDF's associated with chance probabilities of .05 or less were considered and rating items associated with chance probabilities of .05 or less were used to score retained LDF's. Students in the first study were classified on the basis of two LDF's derived in the first study with Bayesian adjustment of probabilities of group membership. The resulting classifications (predicted lecturer type) were compared with known classifications (actual lecturer type). The independence of known and predicted classifications was tested using Chi-Square analysis (Siegel, 1956).

The second phase of the analysis plan consisted of using the LDF's derived in the first study to classify independent groups of students who experienced and rated the same four lectures during the second study. Students in the second study were classified on the basis of the LDF's *without* Bayesian adjustment of probabilities of group membership. Predicted and known classifications in the second study were compared using Chi-Square analysis.

If the rating scores defined by the LDF's are valid, they should result in correct classification of lecturer types a significant proportion of the time. A chance probability of .05 or less (two-tailed test) was established for Type I errors in testing this hypothesis.

Results

Two significant LDF's accounting for 73 and 19 percent of the variance, respectively, were derived in the first study. Six rating variables were associated with significant coefficients. Standardized coefficients associated with significant functions are presented in Table 1. Each significant function was interpreted by considering the rating item associated with the highest coefficient and items associated with coefficients equal to or greater than half that amount (Tatsuoka, 1971). In the case of the first LDF, the results were straightforward, i.e., there was one important coefficient for the rating item pertaining to faculty "enthusiasm."

High positive coefficients for the second LDF were observed for rating items pertaining to "spoke understandably," "broadened my

interest in the subject," and "stressed important material." A high negative coefficient on the second LDF was observed for the rating item pertaining to "gave examples to explain." Thus, a high positive score on the second LDF appears to indicate understandability or clarity of the subject matter or material presented. Given that unrelated examples and details of studies without results were used as filler material in the low information lectures, it is not surprising that a high score on the second LDF was associated with a low rating for "gave examples."

Classifications of lecturer types in the first study, using LDF's derived from the same data are shown in Table 2. Classifications were correct for both faculty characteristics for 77 of 115 students (approximately 67 percent). The classifications of lecturers in terms of enthusiasm (high versus low) were correct for 99 of 115 students (approximately 86 percent). Information-giving differences (high versus low) were correctly classified for 89 of 115 students (approximately 77 percent).

However, the ultimate goal of the current research was to determine the extent to which LDF's that appear to be valid in one study are generalizable to groups of students other than those for whom the function weights were derived. The results of classifications of lecturer types in the second study using the two LDF's derived in the first study are shown in Table 3. Classifications by 26 of the 69 students were correct for both faculty characteristics (approximately a 39 percent hit rate). Although not very impressive, these results indicate that actual and predicted classifications are not independent when marginal totals are considered in computing expected frequencies ($\chi^2 = 8.2$, df = 1, $p < .01$, corrected for continuity).

Differences in the validity of the two LDF's are apparent in Table 3 for the two faculty characteristics manipulated in the second study. Classifications of lecturers in terms of enthusiasm were correct for 53 of 69 students (approximately 77 percent). This relationship represents a validity coefficient of .55 ($p < .001$) when expressed as phi and a Chi-Square value of 21.07 (df = 1, $p < .001$). On the other hand, classifications in terms of information-giving were correct for 34 of 69 students (approximately 49 percent). This degree of association is

represented by a phi coefficient of .07 and a Chi-Square of .30 (df = 1), neither of which is significant.

Some insight into the nature of student errors in classifying faculty according to information-giving can be gained from further analysis of the data presented in Table 3. First, it is helpful to note that 53 students in the second study classified the lecturer they experienced as high in information-giving, whereas, only 30 actually saw a high information lecture. Fourteen of the 23 errors in classification (i.e., approximately 61 percent of the errors) were made by students who actually saw a lecture that was low in information-giving (only four teaching points covered) and that was presented in an *enthusiastic* manner. Thus, the majority of errors in detecting differences in information-giving appear to be the result of bias in student ratings due to differences in faculty enthusiasm. This is an example of what has been termed the "Doctor Fox Effect" (Ware and Williams, 1975).

Finally, some insight into the validity of the LDF's on a *group* basis can be gained by way of an analysis of group centroids. Group centroids are presented in Table 4 and are plotted in Figure 2 for all eight groups.

The two LDF's clearly differentiate faculty presentations with respect to information-giving and enthusiasm in the first study. The centroids for groups of students who experienced high information lectures are high in relation to the second LDF (subject matter). Likewise, centroids for groups of students who experienced enthusiastic lecture presentations in the first study are high in relation to the first LDF (enthusiasm) and groups of students who experienced lectures delivered so as to be low in enthusiasm in the first study are low in relation to the first LDF.

The same trends are apparent for student groups in the second study with respect to the first LDF but not the second LDF (subject matter). Centroids for groups of students who experienced enthusiastic lecture presentations in the second study are high on the first LDF as they should be. However, these groups are not accurately differentiated on the second LDF. The centroids for groups of students who experienced lecture presentations delivered so as to be low in enthusiasm in the

second study are low on the first LDF as they should be and are somewhat differentiated on the second LDF as they should be.

Discussion

The first issue addressed in the current study concerned the validity of student ratings of faculty with respect to differences in faculty presentation manner during lectures. This faculty characteristic has been discussed under a variety of names, including "charisma," "personality," "dynamism," "expressiveness," and "enthusiasm." The use of discrimination analysis to develop student rating scales that are valid with respect to faculty enthusiasm is supported by the current study findings. A stringent cross-validation of an enthusiasm rating scale was successful. It should also be noted that a single rating scale item that correlates highest with this scale is a valid measure of faculty enthusiasm. In other words, many items and complicated weights are not necessary if the correct items are used.

The second issue concerned the validity of student ratings with respect to differences in faculty information-giving during lecture presentations. It is generally accepted that one goal of lecture presentations is the dissemination of information. This may not be the only goal but most agree that information-giving *is* a goal. We had hoped that the use of discriminant analysis and data gathered during controlled experiments (as a means for selecting and weighting items) would improve the sensitivity of student ratings to actual differences in information-giving. In order to be generally useful, such improvements would have to be generalizable, i.e., valid across student groups. Unfortunately, scoring methods that increased the validity of student ratings with respect to differences in faculty information-giving were not valid when used with independent groups of students who experienced the same lecturers.

On the basis of previous Doctor Fox studies it has been established that: a) student ratings are not sensitive to amount of information covered in lectures even though student achievement is affected directly (Ware and Williams, 1975); b) sensitizing students to the content of

lectures by adding an incentive for learning from the lectures does not improve student accuracy in rating content coverage (Williams and Ware, 1976a); c) accuracy of ratings of content coverage does not improve when students are exposed to a second lecture under the same conditions (Williams and Ware, 1976b). In other words, providing students with an additional exposure to the lecturer does not enhance the students' ability to see through the Doctor Fox effect. The present study indicates that empirically based selection and weighting of items does not improve the validity of student rating scores in detecting real differences in information-giving.

However, this study and others provide some leads. In the current study (data not reported), the information-giving rating score used alone was less accurate in detecting actual differences in information-giving than when used in conjunction with the enthusiasm rating scale. Trends in this study and previous "Doctor Fox" studies indicate that differences in faculty information-giving are easier to detect using rating scores when faculty members are less enthusiastic in their lecture presentations. Perhaps a valid information-giving rating procedure can be developed that uses different items and different weights depending on the degree of enthusiasm detected through use of an enthusiasm rating scale.

Improvements in the content of items used in student-faculty rating instruments may also further improve the validity of rating scores. A replication of the current study with additional items pertaining to content and clarity of subject matter indicates promise for improving the validity of rating scores in relation to differences in information-giving.

Finally, it may be possible to provide simple instruction which will enable students to more accurately rate instructor information-giving. This does not appear to have been tried with respect to student-faculty ratings of instruction but a study by Browne and Anderson (1975) suggests that the idea is worth investigation.

Current practices in the evaluation of teaching effectiveness are limited almost entirely to proxy measures, namely student ratings. Direct observation of faculty by fellow specialists and direct measures

of student achievement are almost never used in the evaluation of teaching effectiveness. Students are valid observers of how enthusiastic faculty members are but are not valid observers of two other important aspects of teaching effectiveness, namely, a) differences in lecturer information-giving and b) differences in their own achievement as a result of the lecture.

Until student rating scales are constructed so as to be valid with respect to differences in faculty information-giving, we suggest that the best (if not the only) way to evaluate such differences in faculty is direct observation by trained evaluators and the best (if not only) way to evaluate student achievement is an achievement test in conjunction with proper controls. The "state of the art" is that student ratings of faculty are of little or no use with respect to differences in faculty information-giving and student achievement.

REFERENCES

Browne, R. J., and Anderson, O. R. Lesson Kinetic Structure Analysis as Related to Pupil Awareness and Achievement. *Journal of Educational Psychology*, 1974, *55*, 864-871.

Coats, W. D. and Smidchens, U. Audience Recall as a Function of Speaker Dynamism. *Journal of Educational Psychology*, 1966, *57(4)*, 189-191.

Coffman, W. E. Determining Students' Concepts of Effective Teaching from their Ratings of Instructors. *Journal of Educational Psychology*, 1954, *45*, 277-286.

Isaacson, R. L., McKeachie, W. J., Milholland, J. E., Lin, J. G., Hofeller, J. W., and Zinn, K. L. Dimensions of Student Evaluations of Teaching. *Journal of Educational Psychology*, 1964, *55*, 344-351.

Mastin, V. E. Teacher Enthusiasm. *Journal of Educational Research*, 1963, *56*, 385-386.

Powell, R. W. The Influence of Grading Standards Upon Learning and Student Evaluation of Instruction. Department of Psychology, University of Southern Florida, Tampa, Florida, 33620 (mimeographed paper).

Rosenshine, B. and Furst, N. Research in Teacher Performance Criteria. In B. O. Smith (ed), *Research in Teacher Education*, Englewood Cliffs, New Jersey: Prentice-Hall, 1971.

Solomon, D. Teacher Behavior Dimensions, Course Characteristics, and Student Evaluations of Teachers. *American Educational Research Journal*, 1966, *3*, 35-47.

Tatsouka, M. M. *Multivariate Analysis: Techniques for Educational and Psychological Research.* New York: John Wiley and Sons, 1971.

Ware, J. E. and Williams, R. G. The Doctor Fox Effect: A Study of Lecturer Effectiveness and Ratings of Instruction. *Journal of Medical Education*, 1975, *50*, 149-156.

Williams, R. G. and Ware, J. E. Validity of Student Ratings Under Different Incentive Conditions: A Further Study of the Doctor Fox Effect. *Journal of Educational Psychology*, 1976 (in press).

Williams, R. G. and Ware, J. E. An Extended Visit with Dr. Fox: Validity of Student Ratings After Repeated Exposures to a Lecturer. 1976, (mimeographed paper).

Figure 1. Student-Faculty Rating Items[a]

The Lecturer:

1. Spoke understandably.
2. Knew if students understood him.
3. Showed an interest in students.
4. Increased your appreciation for the subject.
5. In general, taught effectively.
6. Gave several examples to explain complex ideas.
7. Knew his subject matter.
8. Stressed important material.
9. Was an effective lecturer.
10. Has a good sense of humor.
11. Organized and presented subject matter well.
12. Inspired confidence in his knowledge of the subject.
13. Broadened my interest in the subject.
14. Explained the subject clearly.
15. Increased my knowledge of the subject.
16. Stimulated my thinking.
17. Was enthusiastic about the subject.
18. Made learning enjoyable.

---

[a]Items 1-7 were scored using a five-choice response continuum ranging from "Exceptional Performance" to "Improvement Definitely Needed." Items 8-18 were scored from responses to a five-choice response continuum ranging from "I strongly agree with the statement" to "I strongly disagree with the statement."

Table 1

Standardized Weights for Student-Faculty Rating Variables

Item No.	Content[a]	Standardized Weights[b] LDF I	LDF II
1.	Spoke understandably	.040	.336
6.	Gave examples to explain	.076	-.295
8.	Stressed important material	.061	.151
11.	Organized/presented well	-.141	-.018
13.	Broadened my interest	-.147	.222
17.	Was enthusiastic	.533	-.147

[a]Abbreviated item content (see Figure 1 for verbatim items).

[b]Derived in the first study.

Table 2: Comparison of Predicted and Actual Lecturer Types, First Study

	Actual Lecturer Type	f[a]	Predicted Lecturer Type HI HE	HI LE	LI HE	LI LE
I.	High Information (HI) and High Enthusiasm (HE)	26	15[b]	4	7	0
II.	High Information (HI) and Low Enthusiasm (LE)	40	6	26	3	5
III.	Low Information (LI) and High Enthusiasm (HE)	30	6	0	23	1
IV.	Low Information (LI) and Low Enthusiasm (LE)	19	1	4	1	13

[a]Actual number of students who experienced lecturer type, i.e., sum of row entries.

[b]Underlined table entries indicate frequency of correct predictions for *both* lecturer characteristics, i.e., information-giving and enthusiasm.

Table 3. Comparison of Predicted and Actual Lecturer Types, Second Study

			Predicted Lecturer Type			
			HI	HI	LI	LI
	Actual Lecturer Type	f[a]	HE	LE	HE	LE
I.	High Information (HI) and High Enthusiasm (HE)	13	<u>8</u>[b]	1	4	0
II.	High Information (HI) and Low Enthusiasm (LE)	17	4	<u>11</u>	2	0
III.	Low Information (LI) and High Enthusiasm (HE)	21	14	3	<u>4</u>	0
IV.	*Low Information* (LI) and Low Enthusiasm (LE)	18	3	9	3	<u>3</u>

[a] Actual number of students who experienced lecturer type, i.e., sum of row entries.

[b] Underlined table entries indicate frequency of correct predictions for *both* lecturer characteristics, i.e., information-giving and enthusiasm.

Table 4. Means and Standard Deviations for Student-Faculty Rating Function Scores, Two Studies

Study/Lecturer Type	N	LDF I Mean	LDF I Standard Deviation	LDF II Mean	LDF II Standard Deviation
First Study					
I. High Information and High Enthusiasm	26	.329	.230	.187	.326
II. High Information and Low Enthusiasm	40	-.297	.373	.148	.416
III. Low Information and High Enthusiasm	30	.428	.341	-.188	.320
IV. Low Information and Low Enthusiasm	19	-.501	.373	-.270	.284
Second Study					
I. High Information and High Enthusiasm	13	.430	.279	.276	.322
II. High Information and Low Enthusiasm	17	-.074	.384	.215	.289
III. Low Information and High Enthusiasm	21	.397	.269	.214	.442
IV. Low Information and Low Enthusiasm	18	-.394	.494	.099	.485

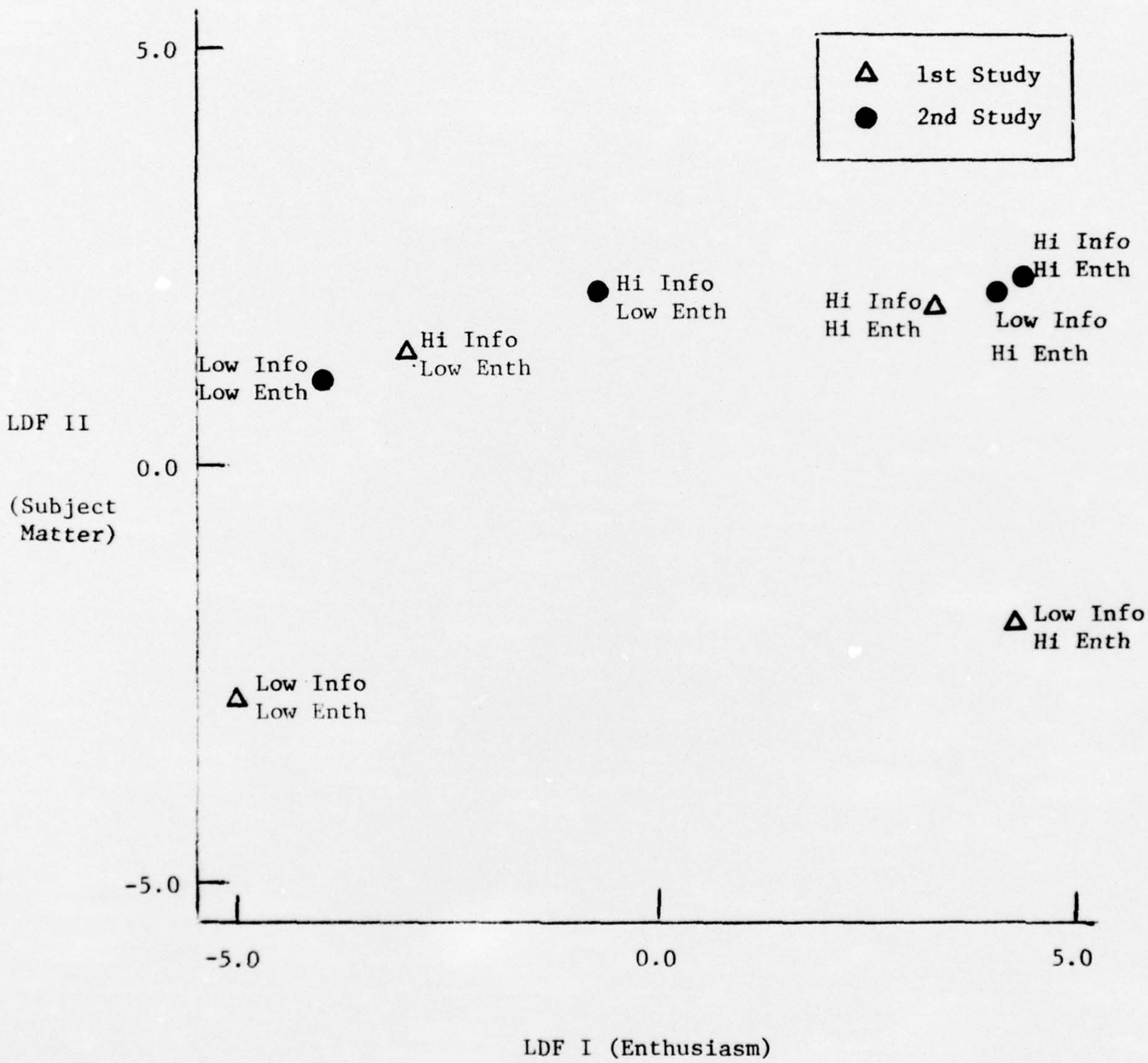

Figure 2

PLOT OF GROUP CENTROIDS FOR STUDENT RATING LDF's